Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense responsive transcription factors in *Populus*

Jin Zhang[1], Yongil Yang[1], Kaijie Zheng[1], Meng Xie[1], Kai Feng[1], Sara S. Jawdy[1], Lee E. Gunter[1], Priya Ranjan[1], Vasanth R. Singan[2], Nancy Engle[1], Erika Lindquist[2], Kerrie Barry[2], Jeremy Schmutz[2,3], Nan Zhao[4], Timothy J. Tschaplinski[1], Jared LeBoldus[5], Gerald A. Tuskan[1], Jin-Gui Chen[1,*] & Wellington Muchero[1,*]

1. Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

2. U.S. Department of Energy Joint Genome Institute, Walnut Creek, California, USA

3. HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA

4. Institute of Agriculture, University of Tennessee, Knoxville, Tennessee, USA

5. Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA

[*] Correspondence should be addressed to J.-G.C. (chenj@ornl.gov) and W.M. (mucherow@ornl.gov).

Summary

We integrated genome-wide associated studies (GWAS) and expression-based quantitative trait loci (eQTL) studies in *Populus trichocarpa* to identify genetic elements controlling abundance of *cis*- and *trans*-3-*O*-caffeoylquinic acid, which are known to be the main contributors to the free radical-scavenging activity. Here, we report that abundances of these metabolites were not only significantly associated with single nucleotide polymorphisms (SNPs) in a *Populus* Hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase (*PtHCT2*), but were also correlated with the expression levels of the same gene based on RNA-Seq analysis targeting leaf tissue. eQTL analysis revealed that *PtHCT2* expression was regulated by putative cis-acting elements, which coincided with GWAS SNP associations, and were also located in the W-box element, a binding site for WRKY transcription factors (TFs). Further analyses in co-expression networks, transcriptional response to infection by the fungal pathogen *Sphaerulina musiva*, and *in vitro* validation of transcriptional regulation suggest that *PtHCT2* is involved in both caffeoylquinic acid biosynthesis as well as defense response, and that its expression is regulated by the defense-responsive WRKY TFs.

Keyword:

Genome-wide association studies (GWAS), Metabolome, Hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase (HCT), WRKY, *Populus trichocarpa*

Introduction

Secondary metabolite biosynthesis is a complex and precise process that is catalyzed by numerous enzymes that fall under complex transcriptional regulatory networks [1]. The identification of key regulators in secondary metabolite biosynthesis remains restricted by low throughput techniques. 3-*O*-caffeoylquinic acid, also known as chlorogenic acid (CGA), is the ester of caffeic acid and (−)-quinic acid and functioning as an intermediate in lignin biosynthesis [2]. It is widely distributed among numerous plant species [3] and acts as an antioxidant in both plants and animals [4]. CGA has been shown to prevent cardiovascular disease and other degenerative, age-related diseases in animals, such as reduce blood pressure, anti-inflammatory, anti-diabetic, anti-carcinogenic, and anti-obesity impacts, etc. [5, 6].

In the phenylpropanoid pathway, hydroxycinnamoyl CoA:shikimate/quinate hydroxycinnamoyl transferase (HCT) catalyzes the conversion of coumaroyl CoA to coumaroyl quinate or coumaroyl shikimate and also the reverse reaction converting caffeoyl quinate or caffeoyl shikimate back to caffeoyl CoA [7]. HCT belongs to the BAHD (The BAHD acyltransferase family was named according to the first letter of each of the first four biochemically characterized enzymes of this family including BEAT, AHCT, HCBT and DAT) family of acyl-CoA-dependent transferases. These transferase can use hydroxycinnamoyl-CoAs as a donor for the transfer reaction and acylating a variety of acceptors [8]. Based on biochemical analysis, the switchgrass *HCT* genes, *PvHCT1a* and *PvHCT2a*, exhibited the expected HCT activity and prefer shikimic acid as an acyl acceptor [9]. CcHCT from globe artichoke could accept 3-hydroxyanthranilate as a substrate [10]. In alfalfa, down-regulation of *p*-coumarate 3-hydroxylase (C3H) and HCT improved fermentable sugar yields without acid pretreatment [11]. Based on the one- and two-dimensional nuclear magnetic resonance (NMR) analyses, a substantial increase in H units as well as a concomitant decrease in G and S units in *C3H* and *HCT* down-regulated alfalfa were observed. ^{13}C NMR analysis estimated that *HCT* down-regulation reduced the methoxyl content by ~73%, which was stronger than *C3H* down-regulation (~55-58%) [12].

In a wide range of plant species, lignin provides a physical barrier against initial ingress of pathogens into plant tissues [13]. Lignin or lignin-like phenolic polymers are induced and rapidly deposited in cell walls in response to both biotic and abiotic stress [14-16]. In many cases, “defense” lignin shown to have elevated levels of H units [17, 18]. Based on quantitative trait loci

(QTL) and genome-wide association mapping studies (GWAS) in maize (*Zea mays*), two key enzymes in lignin biosynthesis, HCT and caffeoyl CoA *O*-methyltransferase (CCoAOMT), were identified adjacent to SNPs that were highly associated with variation in the severity of hypersensitive response (HR) triggered by an intragenic recombinant nucleotide binding leucine-rich-repeat (NLR) disease resistance (*R*) gene Rp1-D21 [19]. Two maize HCT homologs (HCT1806 and HCT4918) physically interact with and suppress the HR conferred by Rp1-D21 but not other autoactive NLRs [20]. In *Arabidopsis* and alfalfa, antisense/RNAi suppression of *HCT* exhibited constitutive activation of defense responses [21, 22]. In addition, many other phenolic compounds synthesized by phenylpropanoid pathway, including phenolic phytoalexins, stilbenes, coumarins, and flavonoids, were also implicated in plant defense [23-26]. For instance, the hormone salicylic acid (SA) that involved in defense signaling is also synthesized through phenylpropanoid pathway in some plant species [27, 28]. Furthermore, the expression of genes encoding monolignol biosynthetic enzymes and corresponding protein levels and enzymatic activities were induced under biotic stress in many plant species [29, 30].

As a class of plant-specific transcription factors (TFs), WRKY has been well recognized for its role in regulating abiotic and biotic stresses [31]. The involvement of WRKY TFs in regulation of a variety of phenolic compounds, including lignin [32-34] has been demonstrated before. Loss of function of *AtWRKY12* in *Arabidopsis* or its ortholog in *M. truncatula* resulted in secondary cell wall thickening in pith cells associated with ectopic deposition of lignin, xylan, and cellulose [34]. Moreover, WRKYs have been shown to control the production of flavanol and tannin compounds. For example, *Arabidopsis* WRKY23 regulates the production of flavanols in an auxin-inducible manner and it has a negative feedback on phytohormone signaling [35].

Based on the previous studies, a total of seven *HCT* members were identified in *Populus*. Among them, *PtHCT1* and *PtHCT6* have been linked to lignin biosynthesis due to their xylem-specific expression profile [36]. Through next-generation sequencing in a natural *Populus nigra* population, *PnHCT1* was identified as an essential enzyme in lignin biosynthesis. *Pn*HCT1 converts *p*-coumaroyl-CoA into *p*-coumaroyl shikimate. The mutant allele trees with homozygous *PnHCT1-Δ73*, which encodes a truncated protein, have a 17-fold increase in H lignin units [37].

In this study, we sought to identify the genetic determinants of *cis*- and *trans*-3-*O*-caffeoylquinic acid leaf abundance, measured using gas chromatography-mass spectrometry (GC-MS) on 739

four-year-old unrelated *P. trichocarpa* genotypes from the Clatskanie, OR field site [38]. Here we describe the characterization of another member of the *HCT* family, *PtHCT2* (Potri.018G105500), in *Populus*. After integrated analyses of the whole-genome re-sequencing, transcriptomic and metabolomics data from a natural population of *P. trichocarpa* to facilitate a high-resolution GWAS, *PtHCT2* was identified as a gene encoding an enzyme associated with biosynthesis of *cis*-3-O-caffeoylquinic acid, *trans*-3-O-caffeoylquinic acid, and a partially identified caffeoyl conjugate metabolite. In addition, *PtHCT2* appears to be involved in defense response via the WRKY transcriptional regulatory pathway.

Results

GWAS results suggest *PtHCT2* is associated with three metabolites

In order to identify key regulators involved in poplar metabolites biosynthesis, we analyzed natural variation in secondary metabolite abundances using gas chromatography-mass spectrometry (GC-MS) on 739 four-year-old unrelated *P. trichocarpa* genotypes from the Clatskanie, OR field site [38]. GWAS performed using a panel of >8.2 Million SNPs and nucleotide insertions and deletions (indels) revealed that *cis*- and *trans*-3-*O*-caffeoylquinic acid as well as a partially identified caffeoyl conjugate metabolite with retention time (RT) 16.61 min and key mass-to-charge (m/z) ratios 219 307 283 were significantly associated with the same interval on chromosome (Chr) 18 of the *Populus* reference genome, with the most significant SNP at Chr18:13235329 for *cis*-3-*O*-caffeoylquinic acid and Chr18:13222746 for *trans*-3-*O*-caffeoylquinic acid and the partially identified caffeoyl conjugate (Fig 1a and S1 Table). Two tandemly-duplicated *HCT* paralogs (Potri.018G105400 and Potri.018G105500) were found within this 12.6 kb interval (Fig 1b).

HCTs in poplar are a multigene family generated by duplication events

In *Populus*, *HCT* belongs to a multi-gene family. Based on previous studies, seven *HCT* genes (*PtHCT1-7*) were identified in the *P. trichocarpa* version 1.1 reference genome [36]. However, in the latest *P. trichocarpa* genome (V3.1), two more *HCT* genes were identified and designated as *PtHCT8* (Potri.005G028400) and *PtHCT9* (Potri.018G105400) (S2 Table). These nine paralogs arose either from the Salicoid whole genome duplication or independent tandem duplications events ("W" and "T" in Fig 1a, respectively). Specifically, *PtHCT2/9*, *PtHCT3/4* and *PtHCT5/7/8*

were generated by tandem duplication events and only *PtHCT1/6* were generated by the whole genome duplication event. We compared the nine *PtHCTs* expression patterns across various tissues using date from the *Populus* Gene Atlas Study (S1 Fig). Overall, the *PtHCTs* in paralogous pairs showed similar expression patterns across 24 samples from six tissues. *PtHCT2/9* were highly expressed in root, *PtHCT1/6* were highly expressed in root and stem and *PtHCT3/4* were highly expressed in leaf and stem. *PtHCT5/7/8* are closely located on Chr5, but only *PtHCT5/8* showed more similarity in both phylogenic relationship and expression pattern (Fig 2a and S1 Fig). Based on the correlation analysis, all four *PtHCT* gene pairs (*1/6*, *2/9*, *3/4* and *5/8*) showed significant positive correlation coefficients (S1 Fig).

To evaluate the differences in regulatory elements in *PtHCTs*, we compared the conserved *cis*-acting elements between the promoter regions of paralogous *PtHCTs*. As shown in S2 Fig, ~84.5% *cis*-acting elements containing promoter regions were conserved in 3 kb upstream of translation start sites (TSS) of paralogous *PtHCT2/9*. While only 4.2% regions were conserved in *PtHCT1/6*. Based on the phylogenetic analysis, three closely-located *PtHCTs*, *PtHCT7* (Potri.005G028000), *PtHCT5* (Potri.005G028100) and *PtHCT8* (Potri.005G028400), were phylogenetically grouped together. When we compared their promoter regions, ~64.7% of the regions were conserved between *PtHCT5* and *PtHCT8*, whereas only ~35.9% and 28.5% were conserved between *PtHCT5/7* and *PtHCT8/7*, respectively (S2 Fig), suggesting *PtHCT7* has diverged with *PtHCT5* and *PtHCT8* in this gene cluster.

Abundance of *cis*-3-O-caffeoylquinic acid, *trans*-3-O-caffeoylquinic acid and a partially characterized metabolite positively correlated with the expression of *PtHCT2*

To provide additional support for this association, we performed RNA-Seq analysis on six-year-old trees from the same Clatskanie field sites. In total 390 leaf and 444 xylem transcriptomes were obtained (including 321 leaf and 429 xylem genotypes from the same genotypes used for leaf metabolite profiling). With these data we first performed correlation analysis between transcript and metabolite abundances for the nine *HCT* paralogs. Interestingly, only *PtHCT2* exhibited significant correlation ($P < 0.001$) with *cis*-, *trans*-3-*O*-caffeoyquinic acid and the partially identified caffeoyl conjugate (RT 16.61 min, m/z 219 307 283) across two independent biological replicates of the leaf transcriptome, with 321 and 202 genotypes, respectively (Fig 2 and S3 Fig). These results suggest that abundances of the three metabolites were not only affected by mutations

at the DNA sequence level but were also affected by the expression levels of *PtHCT2* across the population. A similar analysis with the xylem transcriptome did not show any significant correlation between expression and metabolite abundances (S4 Fig).

eQTL analysis of *PtHCT* family

Based on the above data, we propose that *PtHCT2* is the primary regulator of the three metabolites described above among all *HCTs* in the *Populus* GWAS mapping population. Recently, expression-based quantitative trait loci (eQTL) analyses have been used to identify putative *cis*- and *trans*-regulatory elements underlying variation in gene expression that modulates trait expression [39-41]. To expand on the correlations analysis above, we performed eQTL analysis using transcript abundances as the phenotypic variable in the GWAS analysis using the >8.2 Million SNP/indel panel and normalized transcript counts of *PtHCTs* from 390 leaf and 444 xylem transcriptome datasets. Notably, we identified highly significant associations between *PtHCT2* expression and SNP Chr18:13234933 in leaf and Chr18:13249087 in xylem transcriptomes (Fig 3a and S1 Table). This 14.2 kb interval overlapped with the 12.6 kb region containing SNPs with significant GWAS hits for the three metabolites. This was in spite of the fact that metabolite profiles used for GWAS and leaf and xylem tissue used for eQTL analyses were collected from four- and six-year-old plants under heterogeneous field conditions, respectively (Fig 3d). Interestingly, *PtHCT2* was regulated by the same *cis*-eQTLs in both leaf and xylem; and two SNPs in this region (Chr18:13252615 and Chr18:13252693) affected the core sequences of W-box element ("TGAC" or "GTCA"; Fig 3b,c), which is the transcription factor binding site for WRKY TFs that play major roles in defense response [42-44] and secondary wall formation [33, 34] (S4 Table and S5 Fig).

We also sought to evaluate the level of shared or diverged putative transcriptional regulatory elements for the other eight *HCTs*. Among gene pairs in the *PtHCT* family, *PtHCT3/4* shared the same *cis*-eQTLs in both leaf and xylem, while *PtHCT7/8* shared the same *trans*-eQTLs in leaf transcriptome. In contrast, significant eQTLs of *PtHCT2/9* as well as *PtHCT1/6* were divergent between gene pairs (Fig 3a,b).

Non-synonymous SNPs affect active site of *PtHCT2*

To explore the impact of SNPs located in the *PtHCT2/9* gene pair on protein function, we analyzed protein structures of *PtHCT2* and *PtHCT9*. As shown in Fig 4a, the secondary structures

showed high similarity between the two HCTs. We then performed the structural modeling of *PtHCT2* and *PtHCT9* with I-TASSER [45]. Both *PtHCT2* and *PtHCT9* have a similar structure with model of PDB entry 4g0b [46], except that *PtHCT9* carries an octapeptide tail (MIIAGVEK) in N-terminal (Fig 4b,e).

A total of 438 and 106 SNPs were identified in *PtHCT2* and *PtHCT9* genes, respectively (Table 1 and S3 Table). Among the total of nine *PtHCT* genes, *PtHCT2* showed the most variation across the population. 89.3% (391 of 438) SNPs were located in intronic regions of *PtHCT2* and was significantly higher than that in other *PtHCTs* (27.9~65.1%) (Table 1). For non-synonymous SNPs, a total of 19 and 16 non-synonymous SNPs were identified in *PtHCT2* and *PtHCT9* coding region, respectively. We compared the effects of non-synonymous SNPs between *PtHCT2* and *PtHCT9* (Fig 4b,e) and found that four non-synonymous SNPs affected the protein coding in both PtHCT2 and PtHCT9, i.e. G46V, G75E, V239L and S284F in PtHCT2 corresponding to G54V, G83E, V248L and S293F in PtHCT9, respectively. In addition, some non-synonymous SNPs in PtHCT2 affect the coding amino acid to the type in PtHCT9, and vice versa. For example, T21S, I147L, and V188L in PtHCT2 were predicted to change the coding amino acid to the PtHCT9 model (S19, L155 and L196 in PtHCT9). Similarly, I90T, A205T, N243S and I250T in PtHCT9 corresponding to T82, T197, S234 and T241 in PtHCT2 (Fig 4h and S3 Table).

As an important enzyme involved in multiple metabolism steps, PtHCTs could bind to several ligands through active sites. We then compared the active sites potentially affected by the non-synonymous SNPs in PtHCT2 and PtHCT9. In PtHCT2, H243Y and V328L were active sites for ligand COA and H248Y, S284F and V328L were active sites for ligand WCA (Fig 4c,d). While in PtHCT9, only L170I was identified as an active site for 4KE and L170I and S293F as active sites for COA (Fig 4f,g). Among the two PtHCTs, the same active site (S284F in PtHCT2 and S293F in PtHCT9) was identified.

Co-expression network of *PtHCTs*

In order to provide additional context to the proposed function of *PtHCT2*, we constructed co-expression networks for the nine *HCTs* using 24 *P. trichocarpa* transcriptomic data from different tissues (Phytozome). Subnetworks of *PtHCT2* and *PtHCT9* were relatively independent although they were connected by several hub genes (S6 Fig). *PtHCT3* and *PtHCT7* shared the largest set of co-expressed genes suggesting that these two might be involved in the same biological processes.

Between paralogous pairs, *PtHCT1/6*, which is the only one pair generated by the whole-genome duplication event (Fig 2a), had subnetworks that showed significant divergence (S6 Fig). We then performed GO enrichment analysis to compare the functional differences among these subnetworks. Interestingly, genes co-expressed with *PtHCT2* were significantly enriched for "metabolism" and "defense responses", while genes co-expressed with *PtHCT9* were enriched for carbohydrate related processes (S7 Fig). Further, two *WRKYs* (*PtWRKY38* and *PtWRKY45*) were identified in the *PtHCT2* co-expression network (Fig 5a). The WRKY homologs (*AtWRKY11* and *AtWRKY17*; S8 Fig) in *Arabidopsis* have been previously implicated in basal resistance to pathogen infections [43]. To identify the core TFs controlling the *PtHCT2* sub-network, the enriched *cis*-acting elements of 118 genes co-expressed with *PtHCT2* were analyzed using ELEMENT [47]. Interestingly, the most highly enriched regulatory element in the co-expression network was the WRKY binding site (W-box; S4 Table).

Based on the functional classification, we further classified the genes co-expressed with *PtHCT2*. Among the 188 genes co-expressed with *PtHCT2*, 24 (12.8%) and 22 (11.7%) genes were cell wall-related and defense-related, respectively. Outside of these two major clusters, 17 (9%), 15 (8%) and 14 (7.4%) genes were involved in stress response, transport, and proteolysis processes, respectively (Fig 5a). Noticeable, two *WRKY* transcription factors (TFs), *PtWRKY38* and *PtWRKY45*, were found in the *PtHCT2* co-expression network. These *WRKYs* homologous (*AtWRKY11* and *AtWRKY17*, Fig S6) in *Arabidopsis* have been previously implicated in basal resistance [43].

The genes co-expressed with *PtHCT2* also response to *Sphaerulina musiva*

The observed co-expression with defense-related WRKYs is consistent with previous studies which implicated HCTs in host defense against pathogens via salicylic acid (SA) signaling [21, 22, 48, 49] or by direct physical interaction with other proteins [19, 20]. To provide further evidence supporting a role for *PtHCT2* in defense response, we mined a previous RNA-Seq dataset [50] from two *P. trichocarpa* genotypes infected with *Sphaerulina musiva*, an invasive fungal pathogen in western North America. As shown in Fig 6a, genes in the *PtHCT2* co-expression network, including defense response, stress-related, cell wall-related, transport-related and proteolysis-related genes were significantly induced at 24 h and decreased at 72 h after inoculation in resistance genotype BESC-22, while no significant changes in susceptible BESC-801 during

this stage. Among the nine *PtHCTs*, only *PtHCT2* were up-regulated at 24 h post-inoculation (Fig 6b). Noticeable, many group II and group III *WRKYs* were also significantly upregulated at 24 h (Fig 6c). These included homologs of *AtWRKY11* and *AtWRKY17* which acted as negative regulators of basal resistance to *Pseudomonas syringae* pv. in tomato [43, 51]. During the *S. musiva* susceptibility study, 25 out of 100 *PtWRKYs* were significantly induced in the resistant genotype (40.7%, 33.3% and 30% members in group IIc, IIb and III, respectively). In addition, previous studies showed that most of these *PtWRKYs* (especially in group IIb, IIc, and III) showed significant response to multiple treatments, including salicylic acid (SA), methyl jasmonate (MeJA), *Marssonina brunnea* (Mb), wounding, cold and salinity [52] (S9 Fig).

Transient overexpression of *PtWRKYs* enhanced the expression of *PtHCT2*

To validate the transcriptional regulatory relationships between *PtWRKYs* and *PtHCT2*, we analyzed the expression of *PtHCT2* in response to overexpression of select *PtWRKYs* using a poplar protoplast transient expression system [53]. Based on evidence of induction by *S. musiva* and stress treatments mentioned above, we selected *PtWRKY60* (group IIa), *PtWRKY89* (group III) and *PtWRKY93* (group IIc) (Fig 6 and S9 Fig). In addition, *PtWRKY38* and *PtWRKY45* (both in group IId) were selected based on the *PtHCT2* co-expression analysis (Fig 5a). When the five *PtWRKYs* were transiently overexpressed in poplar protoplasts, expression levels of *PtHCT2* were significantly increased (Fig 5b), suggesting that the expression of *PtHCT2* gene is regulated by WRKYs.

Discussion

As a key component of plant innate immunity, SA plays a central role in systemic-acquired resistance (SAR) [54]. SA is synthesized from chorismite via two alternative pathways, phenylalanine ammonia-lyase (PAL)-dependent phenylpropanoid route and isochorismate synthase (ICS)-dependent route [55]. In the phenylpropanoid pathway, PAL catalyzes the conversion of phenylalanine to cinnamate, and thereby initiates phenylpropanoid metabolism. Subsequently, through cinnamate 4-hydrozylase (C4H), 4-coumarate:coenzyme A ligase (4CL) and the specific branch pathways for the formation of monolignols/lignin, benzoic acids, coumarins, stilbenes and flavonoids/isoflavonoids [26]. From these specific branch pathways,

HCT catalyzes the conversion from coumaroyl CoA to coumaroyl quinate or coumaroyl shikimate and from caffeoyl quinate or caffeoyl shikimate to caffeoyl CoA [7].

Among nine *PtHCTs* identified in our study, two of them (*PtHCT1* and *PtHCT6*) were identified as regulators in lignin biosynthesis based on previous studies [37] and their expression patterns in various tissues and μm-scaled wood-forming zone in poplar (S1 and S10 Fig). The function of other *PtHCT* members remains unclear. Here we provide evidence that another member, *PtHCT2*, is involved in both metabolites (*cis*-3-*O*-caffeoylquinic acid, *trans*-3-*O*-caffeoylquinic acid and an unknown metabolite) biosynthesis and defense response in poplar. Interestingly, not only were SNPs located in *PtHCT2* significant associated with the abundance of three metabolites (*cis*-3-*O*-caffeoylquinic acid, *trans*-3-*O*-caffeoylquinic acid, and an unknown metabolite; Fig 1), the expression of *PtHCT2* was positively correlated with the metabolites' abundance across the *Populus* GWAS mapping population (Fig 2b). The expression patterns of *PtHCT2* in different allele at specific SNP site (Fig 2c) further indicate that its expression level was affected by the SNPs located in the *PtHCT2* gene body.

Based on the physical location, eQTLs are categorized as *cis* or *trans*; i.e. *cis* eQTLs represent a polymorphism physically located near the gene itself. For example, a polymorphism located in the promoter region induce differential expression of the gene [39]. Salvi et al. [40] through positional cloning and association mapping identified a major flowering-time QTL (*Vgt1*) located in 70 kb upstream of an *AP2*-like TF, *Vgt1* functions as a *cis*-acting element and affects the transcript levels of the *AP2*-like TF. In *Arabidopsis*, a QTL study based on the glucosinolate content in a population of 403 Bay × Sha recombinant inbred lines showed that all loci controlling expression variation also affected the accumulation of the resulting metabolites [41]. So, the SNP variation in *PtHCT2* might through regulate the gene expression to affect it mediated regulatory pathway. In addition, non-synonymous SNPs within the gene body could affect the active site of PtHCT2. We compared the 3D structures and the amino acids affected by non-synonymous SNPs in the protein coding region of the paralogous pair *PtHCT2/9* (Fig 2). Noticeably, although *PtHCT2* carried more variation within the gene body (89.3% in intron; Table 1), the active site affected by non-synonymous SNPs showed similar patterns between PtHCT2 and PtHCT9 (S284F in PtHCT2 and S293F in PtHCT9; S3 Table), which implies poplar maintained conserved active site to ensure the fundamental function of PtHCT2 during the evolution.

In addition to metabolites biosynthesis, HCTs have ben also been implicated in host defense against pathogens. During defense response, plants will synthesize a series natural product, which can be categorized into three major groups: phytoalexins, phytoanticipins and signal molecules. Many phenylpropanoids exhibit broad-spectrum antimicrobial activity as preformed "phytoanticipins" or inducible "phytoalexins" [26, 56]. In *Arabidopsis* and alfalfa, down-regulation of *HCT* expression resulted in a dwarf phenotype, elevated SA level, increased *PR* gene expression, and constitutive activation of defense responses [21, 22, 48, 49]. When introduce the *NahG* gene (encodes a salicylate hydroxylase that removes SA) into *HCT*-RNAi plants, the plants restored growth to wild type levels with reduced SA and *PR* transcript levels [21]. These studies provided a link between *HCT* and defense response by SA signaling. In addition, HCT can directly involve in the defense response through physically interaction with other proteins. Maize HCTs (HCT1806 and HCT4918) were shown to physically interact with CCoAOMT2 and Rp1 proteins to form complexes, and suppress Rp1-D21-induced HR [19, 20].

Despite this link, the transcriptional hierarchy leading to HCT response to pathogen infection remains unclear. In this study, we observed that *PtHCT2* was differentially expressed between a resistant and susceptible genotype in response to infection by the fungal pathogen *S. musiva* (Fig 6b). In that regard, it expression pattern was highly correlated with the expression of 10 WRKY TFs from group II or III (Fig 6c). Specifically, three group IIa members, AtWRKY18, AtWRKY40 and AtWRKY60 known to form both homocomplexes and heterocomplexes and interact both physically and functionally in response to different types of microbial pathogens, however, AtWRKY18 plays a more important role than the other two [57]. Four *PtWRKYs* (*28*, *71*, *92* and *93*) were significantly differentially expressed when response to *S. musiva* in poplar, and they clustered together with AtWRKY8 and AtWRKY28 in group IIc (Fig S6). In *Arabidopsis*, *AtWRKY8* plays opposite effects on two pathogens, which is a negative regulator of basal resistance to *P. syringae* and positive regulator to *Botrytis cinereal* [44]. In addition, *AtWRKY8* is also involved in the response of long-distance movement of crucifer-infecting tobacco mosaic virus (TMV-cg) through mediating the crosstalk between ABA and ethylene signaling [58]. PtWRKY38 and PtWRKY45 belong to group IId WRKY, but no specific orthologs were identified in *Arabidopsis* based on the phylogenetic tree (Fig S6). In group IId, several *AtWRKYs* were known involved in defense response. For example, the two group IId WRKYs, *AtWRKY11* and *AtWRKY17*, act as negative regulators of basal resistance to *Pseudomonas syringae* pv. *tomato*

(*Pst*) [43]. Moreover, *AtWRKY11* could work with group III member *AtWRKY70* to serve as regulator in rhizobacterium *Bacillus cereus* AR156-induced systemic resistance to *Pst* DC3000 through activating the JA and SA signaling pathway, respectively [51]. *AtWRKY70* is one of the most represented defense genes. Based on the phylogenetic tree, three group III *PtWRKYs* (*54*, *62* and *89*), which were highly induced by *S. musiva*, were closely clustered with *AtWRKY70* (Fig 6 and S8 Fig). Furthermore, WRKYs regulate the biosynthesis of a variety of phenolic compounds, including lignin [34]. Because lignin is derived from the same phenylpropanoid pathway with other specialized metabolites, the WRKYs regulating lignin biosynthesis or deposition will also affect flux to other phenolic-based metabolites through the phenylpropanoid pathway in directly or indirectly manner [31]. Loss of function of *AtWRKY12* in *Arabidopsis* and its ortholog in *M. truncatula*, SECONDARY WALL THICKENING IN PITH (STP), result in ectopic deposition of lignin, cellulose and xylan, and secondary cell wall thickening in pitch cells [34]. Here, we show that five *PtWRKYs* (*38*, *45*, *60*, *89* and *93*) were induced by *S. musiva* (Fig 6 and S8 Fig) and could also act as activators for *PtHCT2* (Fig 5b).

In summary, *PtHCT2* was identified via GWAS and eQTL analyses as a key regulator for biosynthesis of *cis*-and *trans*-3-*O*-caffeoylquinic acid as well as a partially identified caffeoyl conjugate in the *Populus* GWAS mapping population. eQTL mapping revealed that the *cis*-eQTL is the primary regulatory mechanism of *PtHCT2*. The integrated results from co-expression network analysis, *cis*-acting elements enrichment and response to *S. musiva* suggested the expression of *PtHCT2* is regulated by defense-responsive WRKYs, which was further validated in the poplar protoplast transient expression system. This study provides a new insight to into genetic regulation of three important metabolites and lays a foundation for data-driven characterization of the genetic basis of secondary metabolite biosynthesis in complex perennial plants.

Materials and Methods

Plant materials

Leaf sample for metabolite profiling were collected from the Clatskanie field site in July 2012 and leaf and xylem for RNA-Seq analysis were collected from the same site in July 2014. For each sampling plant materials were immediately frozen on dry ice before processing.

Metabolomic analysis

Freeze-dried leaves were ground to 20 mesh with a micro-Wiley mill and ~25 mg DW was subsequently twice extracted with 2.5 mL 80% ethanol overnight and then combined prior to drying a 0.5 ml aliquot in a nitrogen stream. Sorbitol (75 µL of a 1 mg/mL aqueous solution) was added before extraction as an internal standard to correct for differences in extraction efficiency, subsequent differences in derivatization efficiency and changes in sample volume during heating. Dried extracts were silylated for 1 h at 70°C to generate trimethylsilyl (TMS) derivatives, which were analyzed after 2 days with an Agilent Technologies Inc. (Santa Clara, CA) 5975C inert XL gas chromatograph-mass spectrometer as describes elsewhere [59]. Metabolite peak extraction, identification, and quantification were as described previously [59], and unidentified metabolites were denoted by their retention time as well as key m/z ratios.

RNA-Seq and data analysis

Stored tissue was ground in liquid nitrogen and total RNA was extracted using a combined method including CTAB lysis buffer and a Spectrum Total Plant RNA extraction kit (Sigma). Approximately 100mg of flash frozen ground tissue was incubated in 850ul of CTAB buffer (1.0% β-Mercaptoethanol) at 65°C for 5 minutes, 600 µl chloroform:isoamylalcohol (24:1) was added and samples were spun at full speed for 8 minutes. The supernatant (~730 µl) was removed from the top layer and applied to a filter column provided in the Spectrum kit. RNA was precipitated in 500 µl of 100% ethanol and applied to a Spectrum kit binding column. The protocol provided by the Spectrum kit was followed from that point on and the optional on-column DNase treatment was done to rid the samples of residual genomic DNA. RNA quality and quantity were determined using a Nanodrop Spectrophotometer (Thermo Scientific).

Stranded RNA-Seq library(s) were generated and quantified using qPCR. Sequencing was performed on an Illumina HiSeq 2500 (150mer paired end sequencing). Raw fastq file reads were

399 filtered and trimmed using the JGI QC pipeline. Using BBDuk
400 (https://sourceforge.net/projects/bbmap/), raw reads were evaluated for sequence artifacts by kmer
401 matching (kmer=25) allowing 1 mismatch and detected artifacts were trimmed from the 3' end of
402 the reads. RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality
403 trimming was performed using the phred trimming method set at Q6. Following trimming, reads
404 under the length threshold were removed (minimum length 25 bases or 1/3 of the original read
405 length; whichever was longer). Raw reads from each library were aligned to the *P. trichocarpa*
406 reference genome [60] using TopHat2 [61]. Only reads that mapped uniquely to one locus were
407 counted. FeatureCounts [62] was used to generate raw gene counts. Raw gene counts were used to
408 evaluate the level of correlation between biological replicates, using Pearson's correlation to
409 identify which replicates would be used in the DGE analysis. DESeq2 (v1.2.10) [63] was
410 subsequently used to determine which genes were differentially expressed between pairs of
411 conditions. The parameters used to "call a gene" between conditions was determined at a *P*-value
412 $\leq$0.05.

413 GO enrichment was performed using agriGO (http://bioinfo.cau.edu.cn/agriGO/). For the
414 promoter analysis, the *cis*-elements enrichment in *PtHCT2* co-expression network was analyzed
415 using ELEMENT software [47].

416 **Genome-Wide Association Study (GWAS) and eQTL analyses**

417 Whole genome resequencing, SNP/indel calling and SNPeFF analysis for this 545 individuals of
418 this *Populus* GWAS population was previously described by Evans et al. [64]. In this study, we
419 used the same sequencing and analytical pipelines to incorporate an additional 337 genotypes. The
420 resulting SNP and indel dataset is available at http://bioenergycenter.org/besc/gwas/. To assess
421 genetic control, we used the EMMA algorithm in the EMMAX software with kinship as the
422 correction factor for genetic background effects [65] to compute genotype to phenotype
423 associations using 8.253,066 million SNP variants with minor allele frequencies >0.05 identified
424 from whole-genome resequencing. Metabolite abundances from the GC-MS profiling and
425 normalized FPKM transcript counts were used as phenotypes. A *P*-value threshold of 6.1×10^{-09}
426 (0.05/8,253,066) was used to determine significance based on the Bonferroni correction for
427 multiple testing.

Protein structural modeling

The 3D structures of PtHCT2 and PtHCT9 were built using the Iterative Threading ASSEmbly Refinement (I-TASSER, version 5.1) protein structure modeling toolkit [66]. Structure-based functional annotations and ligand/cofactor predictions of the constructed models were carried out using COFACTOR [67].

Co-expression analysis

FPKM values and co-expression relationships of *PtHCTs* were downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). For the co-expression network, a threshold greater than or equal to 0.85 was applied to the resulting. Cytoscape [68] was used to visualize the resulting network.

For overrepresented *cis*-acting elements identification, 2 kb of upstream sequence relative to the transcription start site of genes in *PtHCT2* co-expression network were analyzed using the ELEMENT program [47]. The significant elements were selected at Benjamini-Hochberg FDR *P*-value < 0.05.

Transient overexpression in poplar protoplast

Protoplasts from *Populus* were isolated and subsequently transfected as previous described [53]. The full-length CDS of five *PtWRKYs* (*38*, *45*, *60*, *89* and *93*) were determined according to the sequence information available at Phytozome. Gene specific primers were designed to amplify the full-length CDS of each *PtWRKY* from *P. trichocarpa* cDNA. Subsequently, the CDS of each *PtWRKYs* was introduced into the pENTR™/D-TOPO vector (Life Technologies). The correct product validated by sequencing was transferred into gateway destination vector driven by 2×35S promoter via LR reaction.

RNA extraction and quantitative RT-PCR (qRT-PCR)

Total RNA from transformed and control poplar protoplast were extracted using the Spectrum™ Plant Total RNA isolation kit (Sigma). Three μg of total RNA were reversely transcribed to cDNA using RevertAid Reverse Transcriptase (Thermo Fisher Scientific). qRT-PCR was performed using Maxima SYBR Green/ROX qPCR Master Mix (Thermo Fisher Scientific). *Populus Ubiquitin* was used as an internal control for normalizing the relative transcript level. All PCR

reactions were done with at least three replicates. The primers used for gene clone and qRT-PCR were listed in S5 Table.

Competing Financial Interests

The authors declare no competing financial interests.

Author Contributions

J.-G.C., W.M., T.J.T. and G.A.T. conceived and designed the experiments. J.Z., Y.Y., K.Z, M.X., S.S.J., L.E.G., T.J.T., N.E., N.Z. and J.L. performed the experiments. J.Z., K.F., V.R.S., E.L., K.B., J.S., J.L., T.J.T. and P.R. analyzed the data. J.Z. drafted the manuscript. J.-G.C., W.M., T.J.T. and G.A.T. revised the manuscript. All authors read and approved the manuscript.

Acknowledgement

This research was supported by the Plant-Microbe Interfaces Scientific Focus Area in the Genomic Science Program, the Office of Biological and Environmental Research in the U.S. Department of Energy (DOE) Office of Science and by the DOE BioEnergy Science Center project. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract Number DE-AC05-00OR22725. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Patra B, Schluttenhofer C, Wu Y, Pattanaik S, Yuan L. Transcriptional regulation of secondary metabolite biosynthesis in plants. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms. 2013;1829(11):1236-47.

2. Boerjan W, Ralph J, Baucher M. Lignin biosynthesis. Annual Review of Plant Biology. 2003;54(1):519-46.

3. Sondheimer E. Chlorogenic acids and related depsides. The Botanical Review. 1964;30(4):667-712.

4. Niggeweg R, Michael AJ, Martin C. Engineering plants with increased levels of the antioxidant chlorogenic acid. Nature Biotechnology. 2004;22(6):746-54.

5. Onakpoya I, Spencer E, Thompson M, Heneghan C. The effect of chlorogenic acid on blood pressure: a systematic review and meta-analysis of randomized clinical trials. Journal of Human Hypertension. 2015;29(2):77-81.

6. Tajik N, Tajik M, Mack I, Enck P. The potential effects of chlorogenic acid, the main phenolic components in coffee, on health: a comprehensive review of the literature. European Journal of Nutrition. 2017:1-30.

7. Hoffmann L, Maury S, Martz F, Geoffroy P, Legrand M. Purification, cloning, and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. Journal of Biological Chemistry. 2003;278(1):95-103.

8. Molina I, Kosma D. Role of HXXXD-motif/BAHD acyltransferases in the biosynthesis of extracellular lipids. Plant cell reports. 2015;34(4):587-601.

9. Escamilla-Treviño LL, Shen H, Hernandez T, Yin Y, Xu Y, Dixon RA. Early lignin pathway enzymes and routes to chlorogenic acid in switchgrass (Panicum virgatum L.). Plant molecular biology. 2014;84(4-5):565-76.

10. Andrea M, Cinzia C, Sergio L, van Beek Teris A, Luca G, Francesco RS, et al. Production of novel antioxidative phenolic amides through heterologous expression of the plant's chlorogenic acid biosynthesis genes in yeast. Metabolic engineering. 2010;12(3):223-32.

11. Chen F, Dixon RA. Lignin modification improves fermentable sugar yields for biofuel production. Nature biotechnology. 2007;25(7):759.

12. Pu Y, Chen F, Ziebell A, Davison BH, Ragauskas AJ. NMR characterization of C3H and HCT down-regulated alfalfa lignin. BioEnergy Research. 2009;2(4):198.

13. Bonello P, Storer AJ, Gordon TR, Wood DL, Heller W. Systemic effects of Heterobasidion annosum on ferulic acid glucoside and lignin of presymptomatic ponderosa pine phloem, and potential effects on bark-beetle-associated fungi. J Chem Ecol. 2003;29(5):1167-82. PubMed PMID: 12857029.

14. Wuyts N, Lognay G, Swennen R, De Waele D. Nematode infection and reproduction in transgenic and mutant Arabidopsis and tobacco with an altered phenylpropanoid metabolism. J Exp Bot. 2006;57(11):2825-35. doi: 10.1093/jxb/erl044. PubMed PMID: 16831845.

15. Menden B, Kohlhoff M, Moerschbacher BM. Wheat cells accumulate a syringyl-rich lignin during the hypersensitive resistance response. Phytochemistry. 2007;68(4):513-20. doi: 10.1016/j.phytochem.2006.11.011. PubMed PMID: 17188312.

16. Sattler SE, Funnell-Harris DL. Modifying lignin to improve bioenergy feedstocks: strengthening the barrier against pathogens? Frontiers in plant science. 2013;4.

17. Robertsen B, Svalheim O. The Nature of Lignin-Like Compounds in Cucumber Hypocotyls Induced by Alpha-1,4-Linked Oligogalacturonides. Physiol Plantarum. 1990;79(3):512-8. PubMed PMID: WOS:A1990DP57200015.

18. Lange BM, Lapierre C, Sandermann H, Jr. Elicitor-Induced Spruce Stress Lignin (Structural Similarity to Early Developmental Lignins). Plant Physiol. 1995;108(3):1277-87. PubMed PMID: 12228544; PubMed Central PMCID: PMCPMC157483.

19. Wang G-F, Balint-Kurti P. Maize homologs of CCoAOMT and HCT, two key enzymes in lignin biosynthesis, form complexes with the NLR Rp1 protein to modulate the defense response. Plant Physiol. 2016:pp. 00224.2016.

20. Wang G-F, He Y, Strauch R, Olukolu BA, Nielsen D, Li X, et al. Maize homologs of hydroxycinnamoyltransferase, a key enzyme in lignin biosynthesis, bind the nucleotide binding leucine-rich repeat Rp1 proteins to modulate the defense response. Plant Physiol. 2015;169(3):2230-43.

21. Gallego-Giraldo L, Escamilla-Trevino L, Jackson LA, Dixon RA. Salicylic acid mediates the reduced growth of lignin down-regulated plants. Proc Natl Acad Sci U S A. 2011;108(51):20814-9. doi: 10.1073/pnas.1117873108. PubMed PMID: WOS:000298289400106.

22. Gallego-Giraldo L, Jikumaru Y, Kamiya Y, Tang YH, Dixon RA. Selective lignin downregulation leads to constitutive defense response expression in alfalfa (*Medicago sativa* L.). New Phytol. 2011;190(3):627-39. doi: 10.1111/j.1469-8137.2010.03621.x. PubMed PMID: WOS:000289641600015.

23. Yu O, Jung WS, Shi J, Croes RA, Fader GM, McGonigle B, et al. Production of the isoflavones genistein and daidzein in non-legume dicot and monocot tissues. Plant Physiol. 2000;124(2):781-93. doi: DOI 10.1104/pp.124.2.781. PubMed PMID: WOS:000089962600030.

24. Lange BM, Lapierre C, Sandermann H. Elicitor-Induced Spruce Stress Lignin - Structural Similarity to Early Developmental Lignins. Plant Physiol. 1995;108(3):1277-87. PubMed PMID: WOS:A1995RJ23500049.

25. Doster MA, Bostock RM. Quantification of Lignin Formation in Almond Bark in Response to Wounding and Infection by Phytophthora Species. Phytopathology. 1988;78(4):473-7. doi: DOI 10.1094/Phyto-78-473. PubMed PMID: WOS:A1988M916100019.

26. Dixon RA, Achnine L, Kota P, Liu CJ, Reddy MSS, Wang LJ. The phenylpropanoid pathway and plant defence - a genomics perspective. Mol Plant Pathol. 2002;3(5):371-90. doi: DOI 10.1046/j.1364-3703.2002.00131.x. PubMed PMID: WOS:000178157500008.

27. Lozovaya VV, Lygin AV, Zernova OV, Ulanov AV, Li S, Hartman GL, et al. Modification of phenolic metabolism in soybean hairy roots through down regulation of chalcone synthase or isoflavone synthase. Planta. 2007;225(3):665-79. doi: 10.1007/s00425-006-0368-z. PubMed PMID: 16924535.

28. Dicko MH, Gruppen H, Barro C, Traore AS, van Berkel WJ, Voragen AG. Impact of phenolic compounds and related enzymes in sorghum varieties for resistance and susceptibility to biotic and abiotic stresses. J Chem Ecol. 2005;31(11):2671-88. doi: 10.1007/s10886-005-7619-5. PubMed PMID: 16273434.

29. Truman W, de Zabala MT, Grant M. Type III effectors orchestrate a complex interplay between transcriptional networks to modify basal defence responses during pathogenesis and resistance. Plant J. 2006;46(1):14-33. doi: 10.1111/j.1365-313X.2006.02672.x. PubMed PMID: WOS:000236035700002.

30. Zhao JW, Buchwaldt L, Rimmer SR, Sharpe A, McGregor L, Bekkaoui D, et al. Patterns of differential gene expression in Brassica napus cultivars infected with Sclerotinia sclerotiorum. Mol Plant Pathol. 2009;10(5):635-49. doi: 10.1111/J.1364-3703.2009.00558.X. PubMed PMID: WOS:000268793200006.

31. Schluttenhofer C, Yuan L. Regulation of specialized metabolism by WRKY transcription factors. Plant Physiol. 2015;167(2):295-306.

32. Naoumkina MA, He X, Dixon RA. Elicitor-induced transcription factors for metabolic reprogramming of secondary metabolism in Medicago truncatula. BMC plant biology. 2008;8(1):132.

33. Guillaumie S, Mzid R, Méchin V, Léon C, Hichri I, Destrac-Irvine A, et al. The grapevine transcription factor WRKY2 influences the lignin pathway and xylem development in tobacco. Plant Molecular Biology. 2010;72(1-2):215.

34. Wang H, Avci U, Nakashima J, Hahn MG, Chen F, Dixon RA. Mutation of WRKY transcription factors initiates pith secondary wall formation and increases stem biomass in dicotyledonous plants. Proc Natl Acad Sci U S A. 2010;107(51):22338-43.

35. Grunewald W, De Smet I, Lewis DR, Löfke C, Jansen L, Goeminne G, et al. Transcription factor WRKY23 assists auxin distribution patterns during Arabidopsis root development through local control on flavonol biosynthesis. Proceedings of the National Academy of Sciences. 2012;109(5):1554-9.

36. Shi R, Sun Y-H, Li Q, Heber S, Sederoff R, Chiang VL. Towards a systems approach for lignin biosynthesis in Populus trichocarpa: transcript abundance and specificity of the monolignol biosynthetic genes. Plant and Cell Physiology. 2009;51(1):144-63.

37. Vanholme B, Cesarino I, Goeminne G, Kim H, Marroni F, Van Acker R, et al. Breeding with rare defective alleles (BRDA): a natural Populus nigra HCT mutant with modified lignin as a case study. New Phytol. 2013;198(3):765-76. doi: 10.1111/nph.12179. PubMed PMID: WOS:000321866600001.

38. Muchero W, Guo J, DiFazio SP, Chen J-G, Ranjan P, Slavov GT, et al. High-resolution genetic mapping of allelic variants associated with cell wall chemistry in *Populus*. BMC Genomics. 2015;16(1):24.

39. Hansen BG, Halkier BA, Kliebenstein DJ. Identifying the molecular basis of QTLs: eQTLs add a new dimension. Trends Plant Sci. 2008;13(2):72-7.

40. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci U S A. 2007;104(27):11376-81. doi: 10.1073/pnas.0704145104. PubMed PMID: 17595297; PubMed Central PMCID: PMCPMC2040906.

41. Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ. Linking metabolic QTLs with network and *cis*-eQTLs controlling biosynthetic pathways. PLoS Genet. 2007;3(9):1687-701. doi: 10.1371/journal.pgen.0030162. PubMed PMID: 17941713; PubMed Central PMCID: PMCPMC1976331.

42. Eulgem T. Dissecting the WRKY web of plant defense regulators. Plos Pathog. 2006;2(11):1028-30. doi: ARTN e126

10.1371/journal.ppat.0020126. PubMed PMID: WOS:000242787100002.

43. Journot-Catalino N, Somssich IE, Roby D, Kroj T. The transcription factors WRKY11 and WRKY17 act as negative regulators of basal resistance in *Arabidopsis thaliana*. The Plant Cell. 2006;18(11):3289-302.

44. Chen L, Zhang L, Yu D. Wounding-induced WRKY8 is involved in basal defense in *Arabidopsis*. Molecular Plant-Microbe Interactions. 2010;23(5):558-65.

45. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res. 2015;43(W1):W174-W81.

46. Lallemand LA, Zubieta C, Lee SG, Wang YC, Acajjaoui S, Timmins J, et al. A Structural Basis for the Biosynthesis of the Major Chlorogenic Acids Found in Coffee. Plant Physiol. 2012;160(1):249-60. doi: 10.1104/pp.112.202051. PubMed PMID: WOS:000308675100024.

47. Mockler T, Michael T, Priest H, Shen R, Sullivan C, Givan S, et al., editors. The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. Cold Spring Harbor Symposia on Quantitative Biology; 2007: Cold Spring Harbor Laboratory Press.

48. Hoffmann L, Besseau S, Geoffroy P, Ritzenthaler C, Meyer D, Lapierre C, et al. Silencing of hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. The Plant Cell. 2004;16(6):1446-65.

49. Li X, Bonawitz ND, Weng J-K, Chapple C. The growth reduction associated with repressed lignin biosynthesis in *Arabidopsis thaliana* is independent of flavonoids. The Plant Cell. 2010;22(5):1620-32.

50. Muchero W, Sondreli KL, Chen J-G, Urbanowicz B, Zhang J, Singan V, et al. Genome wide association mapping reveals loci mediating the interaction between *Sphaerulina musiva* and the forest tree *Populus trichocarpa*. in review. 2017.

51. Jiang C-H, Huang Z-Y, Xie P, Gu C, Li K, Wang D-C, et al. Transcription factors WRKY70 and WRKY11 served as regulators in rhizobacterium *Bacillus cereus* AR156-induced systemic resistance to *Pseudomonas syringae* pv. tomato DC3000 in *Arabidopsis*. Journal of Experimental Botany. 2015;67(1):157-74.

52. Jiang Y, Duan Y, Yin J, Ye S, Zhu J, Zhang F, et al. Genome-wide identification and characterization of the *Populus WRKY* transcription factor family and analysis of their expression in response to biotic and abiotic stresses. Journal of Experimental Botany. 2014;65(22):6629-44.

53. Guo J, Morrell-Falvey JL, Labbé JL, Muchero W, Kalluri UC, Tuskan GA, et al. Highly efficient isolation of *Populus* mesophyll protoplasts and its application in transient expression assays. PLoS One. 2012;7(9):e44908.

54. Vlot AC, Klessig DF, Park SW. Systemic acquired resistance: the elusive signal(s). Curr Opin Plant Biol. 2008;11(4):436-42. doi: 10.1016/j.pbi.2008.05.003. PubMed PMID: WOS:000258852300012.

55. Tsai CJ, Harding SA, Tschaplinski TJ, Lindroth RL, Yuan Y. Genome‐wide analysis of the structural genes regulating defense phenylpropanoid metabolism in Populus. New Phytol. 2006;172(1):47-62.

56. VanEtten HD, Mansfield JW, Bailey JA, Farmer EE. Two Classes of Plant Antibiotics: Phytoalexins versus" Phytoanticipins". The Plant Cell. 1994;6(9):1191.

57. Xu X, Chen C, Fan B, Chen Z. Physical and functional interactions between pathogen-induced Arabidopsis WRKY18, WRKY40, and WRKY60 transcription factors. The Plant Cell. 2006;18(5):1310-26.

58. Chen L, Zhang L, Li D, Wang F, Yu D. WRKY8 transcription factor functions in the TMV-cg defense response by mediating both abscisic acid and ethylene signaling in Arabidopsis. Proceedings of the National Academy of Sciences. 2013;110(21):E1963-E71.

59. Tschaplinski TJ, Standaert RF, Engle NL, Martin MZ, Sangha AK, Parks JM, et al. Down-regulation of the caffeic acid *O*-methyltransferase gene in switchgrass reveals a novel monolignol analog. Biotechnology for biofuels. 2012;5(1):71.

60. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313(5793):1596-604.

61. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013;14(4):R36.

62. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30(7):923-30.

63. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.

64. Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, et al. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. Nature Genetics. 2014;46(10):1089-96.

65. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012;44(7):821-4.

66. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nature methods. 2015;12(1):7-8.

67. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Res. 2017.

68. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2010;27(3):431-2.

Tables

Table 1. SNPs identified in *PtHCT* genes.

SNP Effects	*PtHCTs*								
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
Synonymous coding	20	18	12	14	26	20	21	13	14
Non-synonymous coding*	23	19	32	34	56	18	39	25	16
Start gained*	4	2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Stop gained*	n.a.	n.a.	n.a.	2	5	1	2	n.a.	n.a.
Synonymous stop	n.a.	n.a.	n.a.	n.a.	n.a.	1	n.a.	n.a.	n.a.
Frame shift*	1	1	2	3	1	n.a.	1	1	1
Codon change plus codon insertion*	1	n.a.	n.a.	n.a.	n.a.	n.a.	1	n.a.	1
Splice site acceptor*	n.a.	n.a.	n.a.	n.a.	n.a.	1	n.a.	n.a.	n.a.
Splice site donor*	1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Intron	151	391	73	26	34	100	41	32	62
5'-UTR prime	10	5	2	2	n.a.	7	2	n.a.	n.a.
3'-UTR prime	21	2	13	5	n.a.	42	1	1	12
Summary	**232**	**438**	**134**	**86**	**122**	**190**	**108**	**72**	**106**

Notes: *, functional effects. n.a., not available in this dataset. Details of the non-synonymous SNPs in *PtHCT2* and *PtHCT9* was shown in S3 Table.

Figure Legends

Fig 1. Genome-wide association analysis of three metabolites (*cis*-3-*O*-caffeoylquinic acid, *trans*-3-*O*-caffeoylquinic acid and a partially-identified caffeoyl conjugate; RT 16.61 min, key m/z 219 307 283) accumulation in leaves among the *P. trichocarpa* natural population.

(a) Manhattan plots of the three metabolites. Chromosome (Chr) 18 with highly association with the three metabolites was labelled with green. The location of nine *PtHCT* genes on *Populus* genome was labelled at bottom. The letters "T" and "W" on the links indicate putative tandem duplication and whole-genome duplication, respectively.

(b) Zoom in of Manhattan plots on Chr 18 (upper) and the highly-associated region (yellow background, lower). The highest-associated SNPs located in the gene body of *PtHCT2*.

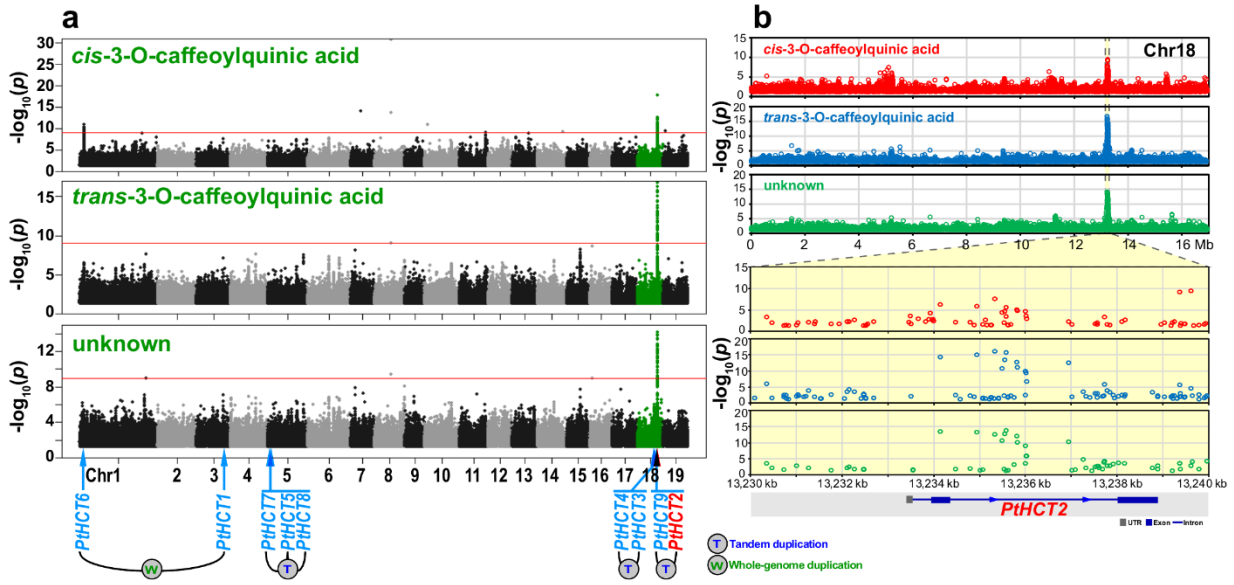

Fig 2. Expression of *PtHCT2* was positively correlated with accumulation of the three metabolites.

(a) Phylogenetic relationship of nine *PtHCTs* in *Populus* genome. Phylogenetic tree was constructed using the Neighbour-Joining methods with 1,000 bootstrap replicates. The letters "T" and "W" on the branches indicate putative tandem duplication and whole-genome duplication, respectively.

(b) The correlation coefficient between gene expression of nine *PtHCTs* and abundance of the three metabolites in leaves across populations from two replicates for independent metabolomic analysis (321 and 202 leaf samples, respectively) of the Clatskanie field site.

(c) Relationships between expression of *PtHCT2*, abundance of the three metabolites and SNPs. Two selected SNPs (Chr18:13235329 and Chr18:13235575) are shown.

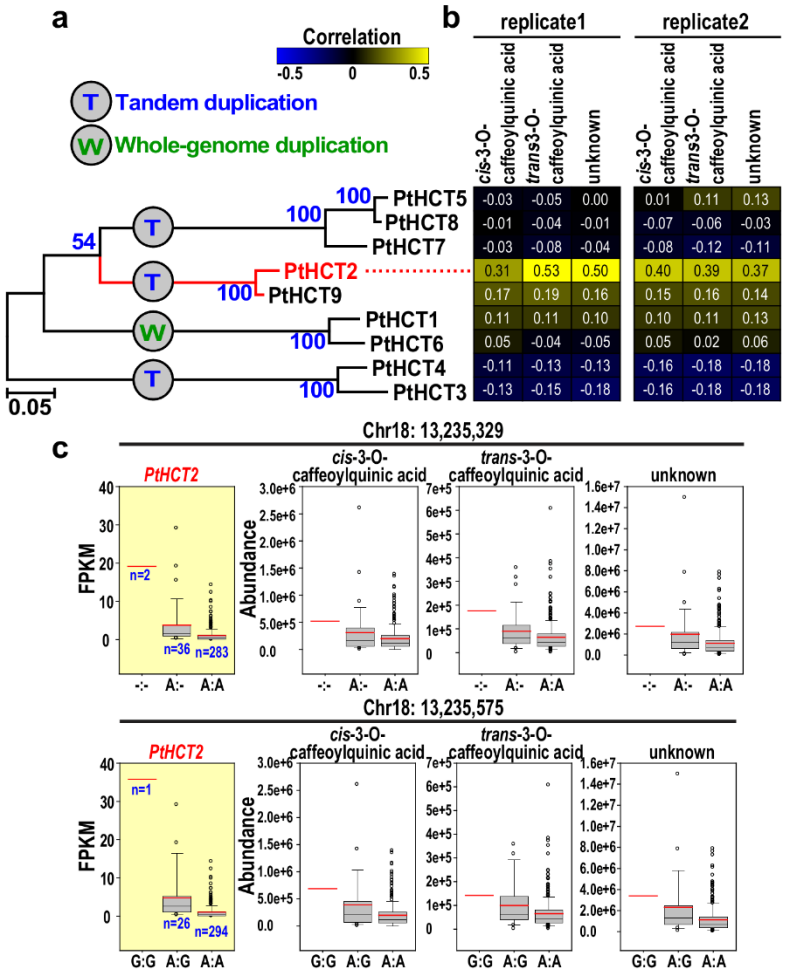

Fig 3. eQTL mapping of *PtHCT* genes in leaf and xylem.

(a) eQTLs associated with nine *PtHCTs* expression in leaf (left panel) and xylem (right panel). Red dots are significant eQTLs with -log10 P value > 5. Blue and green arrows indicate extremely highly associated ($-\log_{10}P$ >10) *cis*- and *trans*-eQTLs, respectively.

(b) Overlapped eQTLs between *PtHCT* gene pairs in leaf and xylem tissues.

(c) *cis*-eQTLs of *PtHCT2*. Among eight overlapped eQTLs of *PtHCT2* between leaf and xylem, six are *cis*-eQTLs, two of which (Chr18:13246177 and Chr18:13252693) affect the core sequences ("GTCA" or "TGAC") of W-box element.

(d) Overlap of interval of *PtHCT2 cis*-eQTL and significant SNP interval of GWAS from the three metabolites.

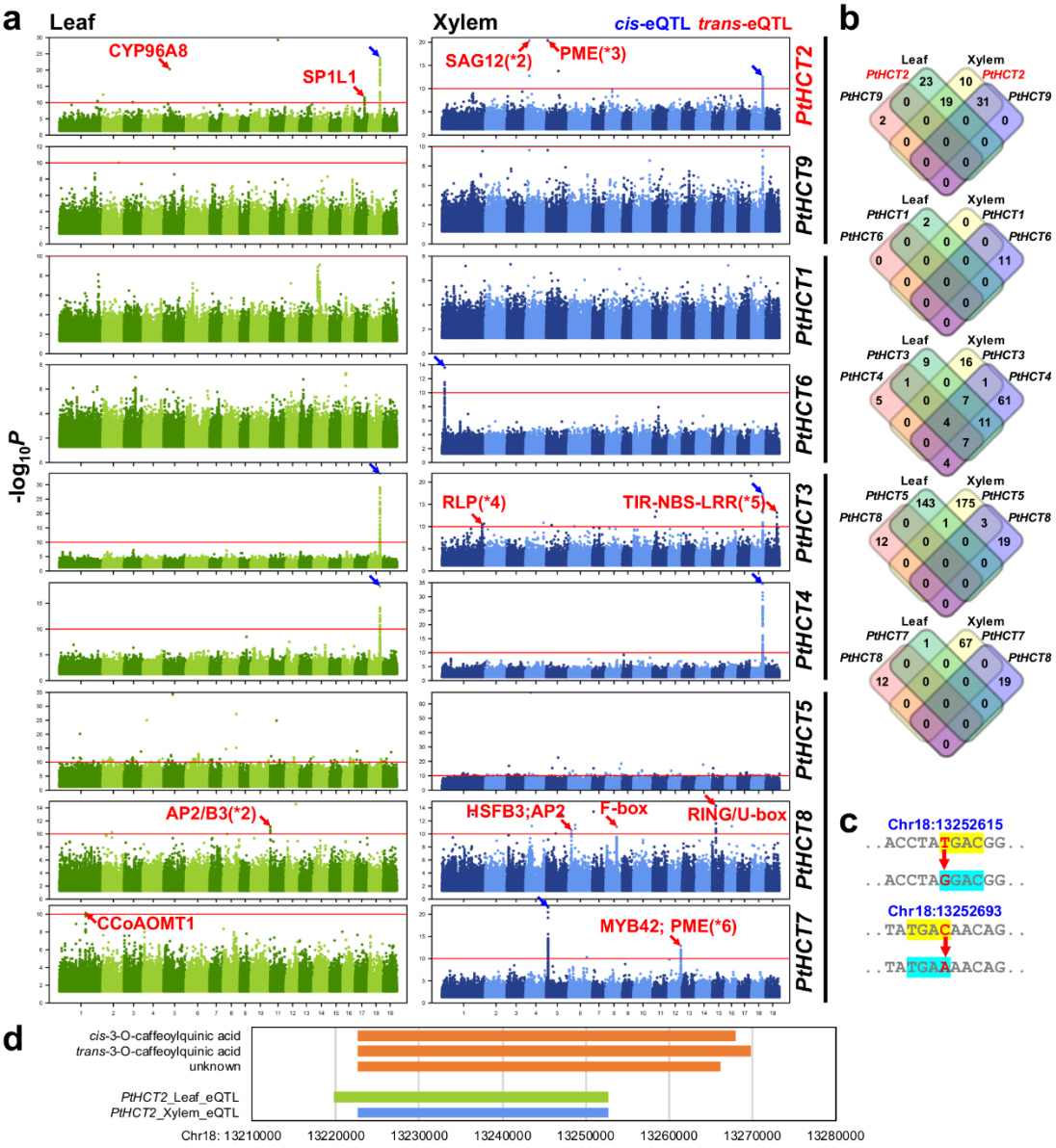

Fig 4. Structural models of PtHCT2 and PtHCT9.

(a) Secondary structures of PtHCT2 and PtHCT9.

(b, e) 3D structures of PtHCT2 and PtHCT9. Yellow chains indicate the PtHCT2 (b) and PtHCT9 (e), blue chains indicate the best identified structural analogs 4g0bA in PDB. Amino acid changes caused by non-synonymous SNPs are labelled in white letters.

(c, d) The active site affected by non-synonymous SNPs in PtHCT2 (H248Y, S284F and V328L).

(f, g) The active site affected by non-synonymous SNPs in PtHCT9 (L170I and S293F).

(h) Sequence alignment of amino acids of PtHCT2 and PtHCT9. Orange shadows, different sequences between PtHCT2 and PtHCT9; green letters, active site affected by non-synonymous SNPs; blue letters, other site affected by non-synonymous SNPs.

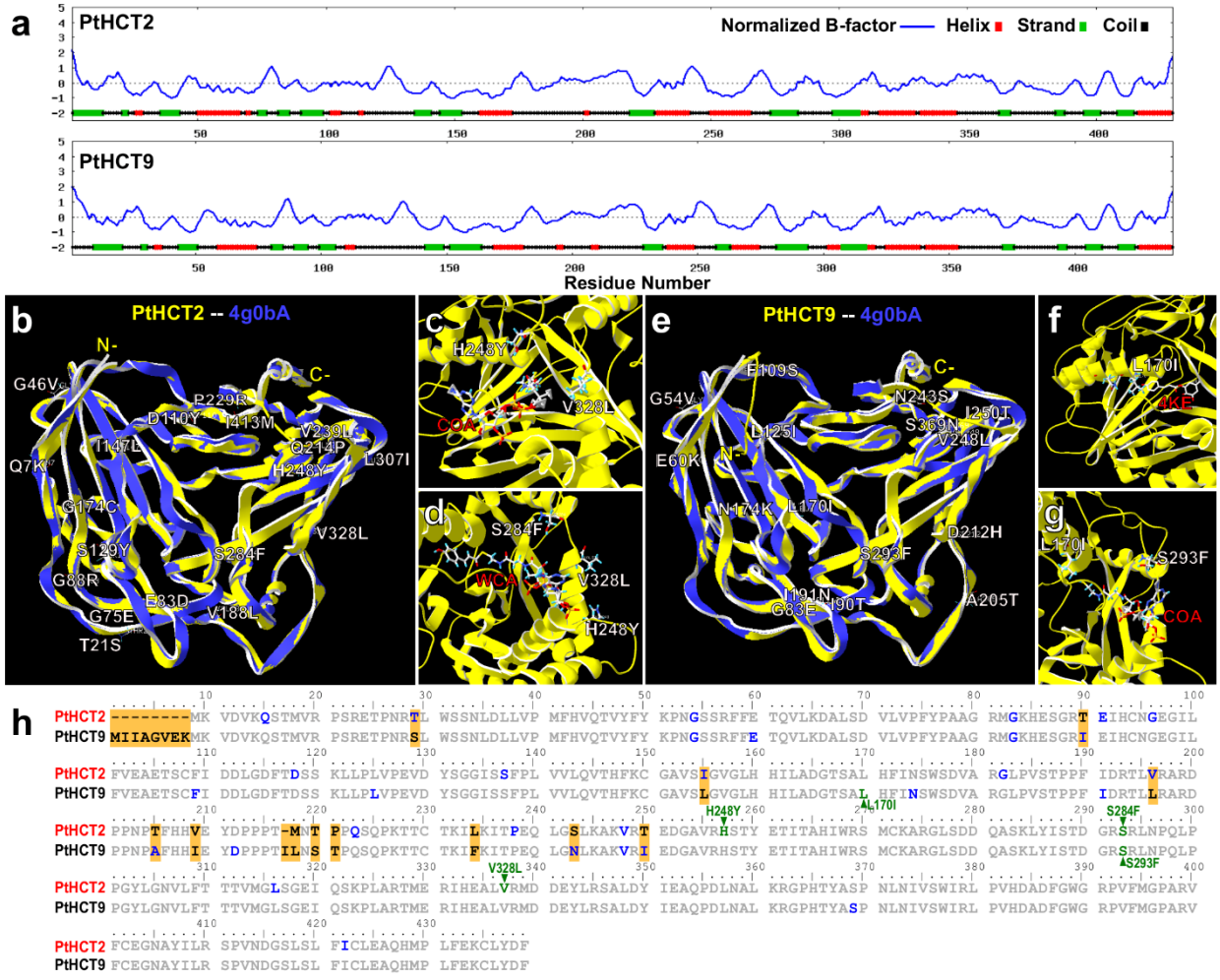

Fig 5. Co-expression network of *PtHCTs* in *Populus*.

(a) Based on the functional annotation, the genes in the *PtHCT2* co-expression network were classified into the following groups: defense response (orange), stress response (pink), cell wall related (green), transport (cyan), proteolysis (purple) and others (grey). Two *WRKYs* (*WRKY38* and *WRKY45*) are among the *PtHCT2* co-expression network.

(b) Regulation of *PtHCT2* by PtWRKYs. Five *PtWRKYs* (*PtWRKY38*, *45*, *60*, *89* and *93*) were transiently overexpressed in Populus protoplasts. The transcript levels of *PtHCT2* were analyzed by using qRT-PCR with three biological replicates.

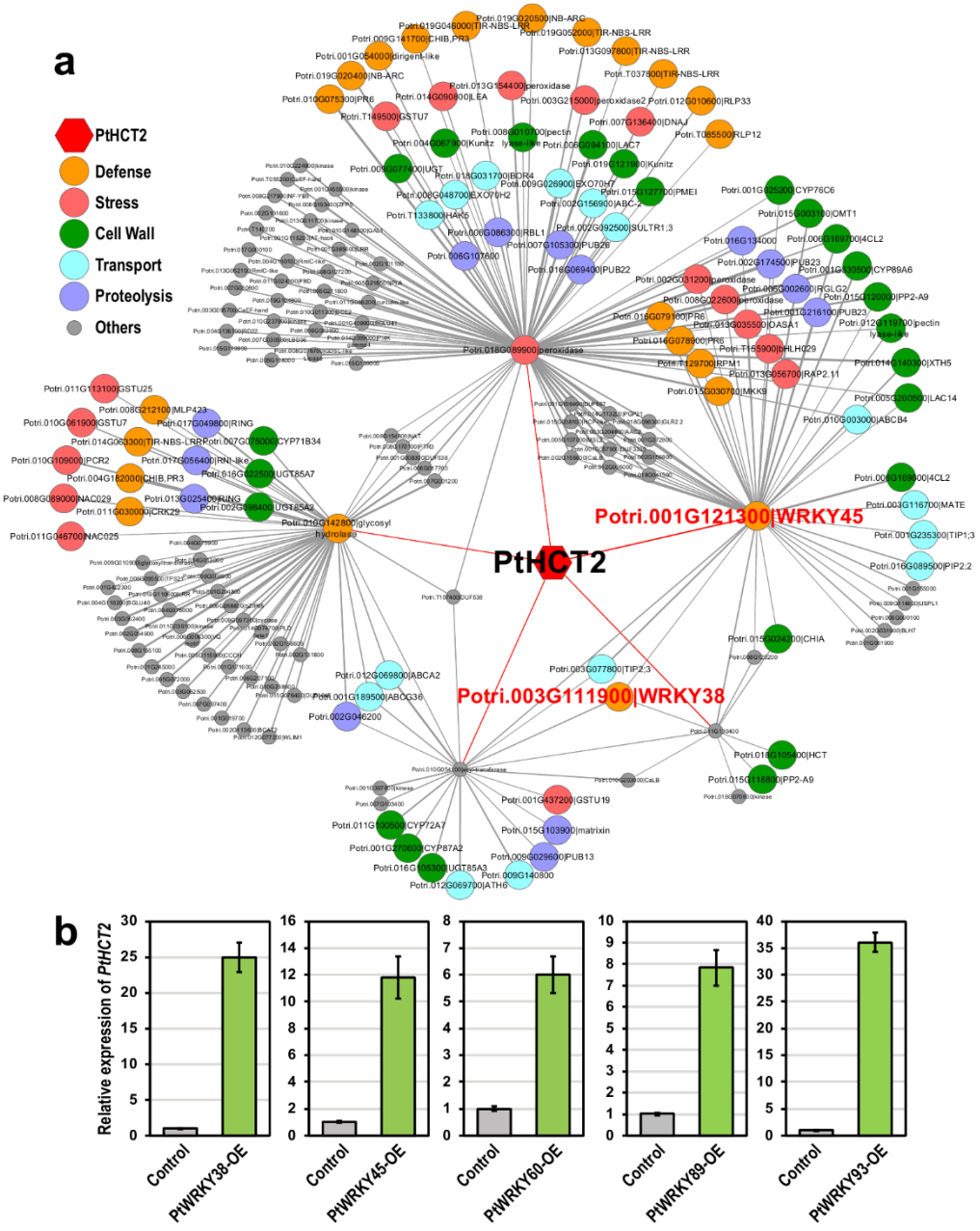

Fig 6. Involvement of genes co-expressed with *PtHCT2* in defense response.

(a) Expression response of five classes genes in *PtHCT2* co-expression network in two *P. trichocarpa* genotypes (BESC22 and BESC801) inoculated with *S. musiva*.

(b) Expression patterns of nine *PtHCTs* response to *S. musiva*. *PtHCT5* was not detected during this process.

(c) Expression patterns of ten selected *PtWRKYs* response to *S. musiva*.